Slide 1 of 31

Title Slide: International Activities Program Sample Designs, Weights, Variance, IRT Scaling, and Plausible Values

Slide 2 of 31

This module provides information about how to appropriately use weights, variance estimation procedures, and plausible values with data from the Progress in International Reading Literacy Study (PIRLS), Trends in International Mathematics and Science Study (TIMSS), and Program for International Student Assessment (PISA). When analyzing these data, the procedures described within this module must be used to assure that the results are representative of the target population and that hypothesis tests are accurate. This means that weights must be applied, standard errors must be correctly calculated, and plausible values must be used when analyzing the test scores. To view the common module where sampling weights and procedures for calculating standard errors in NCES studies are also discussed, click on the underlined screen text 'standard errors'.

First let's review the sample designs for PIRLS, TIMSS, and PISA.

Slide 3 of 31

As was discussed in Common Module 3**,** the data collected from large-scale education studies like PIRLS, TIMSS, and PISA are collected using a complex sample design rather than taking a simple random sample of the target population. Specifically, the samples are selected using a two-stage stratified cluster design.

For the first sampling stage, schools are sampled with probabilities proportional to their size (PPS) from the list of all schools in the population that contain eligible students. Larger schools will therefore have a higher probability of selection than smaller schools, but students in larger schools have a smaller within-school probability of being selected than students in small schools. The probability of a school to be selected is equal to the ratio of the school size (number of students) multiplied by the number of schools to be sampled and divided by the total number of students in the population.

The schools in this list (or sampling frame) may be stratified (or sorted) according to important demographic variables to help ensure proportionate representation in the sample and enhance the precision of the survey results. Many participating education systems employ explicit stratification, where the complete school sampling frame is divided into smaller sampling frames according to some criterion, such as geographic region, to ensure a predetermined number of schools sampled for each stratum. For example, the U.S. sampling frame in PIRLS and TIMSS 2011 was explicitly stratified by three categorical stratification variables: percentage of students eligible for free or reduced-price lunch, type of school (i.e., public or private), and region of the country (i.e., Northeast, Central, West, and Southeast). Stratification also can be done implicitly, a procedure by which schools in a sampling frame are sorted according to a set of stratification variables prior to sampling. For example, the U.S. sampling in PIRLS and

TIMSS 2011 was also implicitly stratified by two categorical stratification variables: community type and minority status.

Although every effort is made to secure the participation of all sampled and eligible schools, substitute schools can be used after all attempts to persuade a sampled school to participate have failed. To avoid sample size losses, the sampling plan identifies, a priori, specific substitute schools for each sampled school that share similar characteristics (such as school size) to the original sample school it may replace.

Slide 4 of 31

In the second stage of sampling, for PIRLS and TIMSS, classrooms of students from the target grade are sampled from within the sampled schools. Intact classrooms are selected because these studies are concerned with what happens within classrooms and schools. For PISA's second stage of sampling, 15-year-old students are selected with equal probability from within the sampled schools. Classrooms are not selected because 15-year-olds are distributed across grades and classes.

More information about the PIRLS and TIMSS sample designs and sampling can be found in *Methods and Procedures in TIMSS and PIRLS 2011,* which is accessible by clicking the underlined screen text, 'PIRLS and TIMSS'. More information about PISA's sample design and sampling can be found in the *PISA Technical Report*, which is accessible by clicking on the underlined screen text 'PISA.'

Slide 5 of 31

While PIRLS, TIMSS, and PISA strive to be as inclusive as possible, exclusions during sampling can occur at the school level, with entire schools being excluded, or within schools, with specific students or entire classrooms excluded. Exclusions should not exceed 5 percent of the target population, and they can only occur for specific reasons. For example, schools that are in remote regions or extremely small schools can be excluded. Students can be excluded if they have intellectual disabilities, functional disabilities, or have insufficient language skills. Accommodations are not provided in PIRLS, TIMSS, and PISA for students with disabilities or students who are unable to read or speak the language of the test. Details on the criteria for exclusions can be found in the studies' school sampling manuals and technical documentation.

Shown in the table are the U.S. exclusion rates. Because 7 percent of U.S. students were excluded in PIRLS and TIMSS in 2011, U.S. results in data tables and figures in international and NCES reports are annotated to indicate that coverage of the defined student population was less than the International Association for the Evaluation of Educational Achievement (or IEA) standard of 95 percent.

Slide 6 of 31

In order to minimize the potential for response biases, PIRLS, TIMSS, and PISA have developed participation or response rate standards that apply to all participating education systems and govern whether or not an education system's data are included in the international datasets and the way in which national statistics are presented in international and NCES reports. Standards for response rates have been set separately at the student and school levels, both with and without the use of substitute schools at the school level. In PIRLS and TIMSS, the minimum acceptable participation rates are 85 percent of both schools and students, or a combined rate (that is, the product of school and student participation) of 75 percent. Education systems not meeting these standards get annotated in the international and NCES reports or could be excluded from these reports and the international datasets. In PISA, the school response-rate target is 85 percent. In order for an education system's data to be included in the international database and the international and NCES reports, a minimum participation rate of 65 percent of schools from the original sample of schools is required. Substitute schools can be used to increase the response rate once the 65 percent benchmark is reached. PISA also requires a minimum student participation rate of 80 percent. As shown in the table, the United States has met participation-rate standards in PIRLS 2011, TIMSS 2011, and PISA 2012.

In addition to participation and response rate standards established by IEA for PIRLS and TIMSS and the Organization for Economic Cooperation and Development (OECD) for PISA, NCES standards require that a nonresponse bias analysis be conducted at any stage of data collection with a weighted unit response rate less than 85 percent.

Slide 7 of 31

PIRLS and TIMSS guidelines call for a minimum of 150 schools to be sampled, with a minimum of 4,000 students assessed. The basic sample design of one classroom per target grade per school is designed to yield a total sample of approximately 4,500 students per population. Education systems with small class sizes or less than 30 students per school are directed to consider sampling more schools, more classrooms per school, or both, to meet the minimum target of 4,000 tested students. In PISA, a minimum of 4,500 students from a minimum of 150 schools is required in each participating education system. The international guidelines specify that within schools, a sample of 35 students is to be selected in an equal probability sample unless fewer than 35 students age 15 are available (in which case all students are selected).

Shown are the U.S. sample sizes—specifically the number of participating schools and the number of assessed students—for PIRLS 2011, TIMSS 2011, and PISA 2012.

Slide 8 of 31

Sampling weights must be used so that estimates produced from the data, such as percentages or average scores, are correct, that is, representative of the target

population. For example, when using U.S. PISA data, weights must be used to obtain results that generalize to U.S. 15-year-old students.

Weights that take into account characteristics of the sample and the selection procedure must be used in analyses of the data. The sampled participants in PIRLS, TIMSS, and PISA did not all have the same chance of being selected into the study. The weights account for these differential probabilities of selection. Additionally, the weights are adjusted to account for school or student nonresponse—to assure that the data are still representative of the population even though some schools and students do not participate.

Slide 9 of 31

The school weight is, essentially, the inverse of the probability of a school being sampled in the first stage of the sampling design. A school-level nonresponse adjustment is applied to compensate for any sampled schools that did not participate and were not replaced. Sampled schools that are found to be ineligible are removed from the calculation of this adjustment. This adjustment is calculated independently for each explicit stratum.

Slide 10 of 31

In PIRLS and TIMSS, classroom weights reflect the probability of the sampled classroom, or classrooms, being selected from among all the classrooms in the school at the target grade level. All classrooms are sampled with equal probability, and the classroom weight is calculated independently for each participating school. The basic class-within-school weight for a sampled class is the inverse of the probability of the class being selected from all of the classes in its school. If a sampled classroom in a school does not participate, or if the participation rate among students in a classroom falls below 50 percent, a classroom-level participation adjustment is made to the classroom weight.

As PISA samples by age (that is, 15-year-old students) rather than grade, classroom weights are not applicable.

Slide 11 of 31

In a simple random sample, students are selected with equal probability of selection, and each student is representing the same number of students in the population. In studies like PIRLS, TIMSS, and PISA that employ a complex sample design, students are selected with unequal probabilities of selection. Thus, students are assigned sampling weights to account for over- or under-representation of particular groups in the final sample. The weight assigned to a student is the inverse of the probability that the student is selected for the sample. When students are weighted, none are discarded, and each contributes to the results for the total number of students represented by the individual students assessed. Some students have a higher weight value than others, meaning that they are representing more students in the target population than sampled students with a lower weight value. The use of sampling weights is necessary for the

computation of sound, nationally representative estimates. Weighting also accounts for student nonresponse, because data cannot be assumed to be randomly missing.

Slide 12 of 31

In PIRLS and TIMSS, the overall student sampling weight is referred to as the total student weight. It is the product of the final weighting components for schools, classrooms, and students, including the nonparticipation adjustments. In PISA, the overall student sampling weight is referred to as the final student weight. It is the product of the final weighting components for schools and students, including the nonparticipation adjustments. If you sum the values of the overall student sampling weights for all participants in an education system, this will approximately equal the size of the target population in that education system.

Slide 13 of 31

PIRLS and TIMSS do not provide for nationally representative samples of teachers. Rather, the participating teachers are those who teach nationally representative samples of PIRLS or TIMSS students. As a consequence, analyses involving teacher data have to be viewed as student-level analyses. Accordingly, the teacher weights are derived from the total student weight, which sums to the national population of study eligible students. PIRLS and TIMSS both include an overall teacher weight, and TIMSS also includes a mathematics teacher weight and a science teacher weight. Teacher weights are calculated by dividing the sampling weight for a student by the number of teachers that the student has. In TIMSS, separate mathematics and science teacher weights are developed by dividing the student sampling weight by the number of mathematics teachers and the number of science teachers that the student has, respectively.

A teacher questionnaire was administered in PISA starting with the 2015 administration. Details about teacher weights will be available when PISA 2015 data get released at the end of 2016.

Slide 14 of 31

The following table summarizes the types of sampling weights that are available within the IAP studies. PIRLS and TIMSS calculate similar weights. As a result of sampling intact classrooms and collecting teacher data, PIRLS and TIMSS calculate more weights than in PISA. Teacher weights will be available in PISA starting with the release of PISA 2015 data.

Slide 15 of 31

In PISA, the final student weight is most commonly used in statistical analyses. A final school weight is also included in the data files, but it should be used with caution given that PISA samples 15-year-old students, not schools. The school questionnaire provides contextual data for nationally representative samples of 15-year-old students.

In PIRLS and TIMSS, the total student weight is commonly used in student-level statistical analyses, though the senate and house weights also are available for student-level analyses with student, home, or school data. Although the total student weight has desirable properties, it also has drawbacks for some analyses. Because the total student weight sums to the student population size in each country, analyses that use this weight and combine countries will have proportionately more students from larger countries and fewer from smaller countries, which may not be desirable for some purposes. Additionally, because the total student weight inflates sample sizes to estimate the population size, using the actual sample size to compute significance tests will give misleading results for analyses weighted by the total student weight. Thus, for cross-country analyses in which countries should be treated equally, the senate weight can be used, which is a transformation of the total student weight that results in a weighted sample size of 500 in each country. The house weight, which is another transformation of the total student weight, ensures that the weighted sample corresponds to the actual sample size in each country.

When using teacher data in student-level analyses, a teacher weight is used. In TIMSS, the mathematics teacher weight will be used when analyzing mathematics teacher data and the science teacher weight will be used when analyzing science teacher data.

The school weight is designed for use in school-level analyses where the schools are the units of analysis. This weight should be used with caution. PIRLS is designed to produce nationally representative estimates of fourth-graders, TIMSS for fourth- and eighth-graders, and PISA for 15-year-old students. So, schools per se are not the target population in any of these studies.

As discussed in the module titled, ‘Considerations for Analysis of IAP Data,’ the IEA IDB Analyzer is a special statistical software package that automatically selects the appropriate weight variable for analysis based on the file types included in the merged data file.

Slide 16 of 31

To illustrate the importance of using weights, consider this example.

A population consists of two strata: stratum 2 is 10 times larger than stratum 1, but we could select the same size sample from each and then use sampling weights. We would assign each student in stratum 1 a weight of 10, and each student in stratum 2 would be assigned a weight of 100.

Slide 17 of 31

Now, let’s say that students in stratum 1 have a mean score of 500, and students in stratum 2 have a mean score of 600. If we ignore the weights, we will compute an unweighted overall mean of 550 because both strata contribute equally to the overall mean. Shown is the formula for computing the unweighted mean as well as the actual calculation.

Slide 18 of 31

However, if we apply the sampling weights to this simple analysis, we will compute an overall mean score of 591. Stratum 2, being larger, contributes more to the overall mean. Shown is the formula for computing the weighted mean as well as the actual calculation.

Slide 19 of 31

Now let's consider another example, this time using U.S. data from PIRLS 2011. As discussed, sampling weights need to be used to obtain results that generalize to the target population. In the above example, notice how the unweighted percentages (shown in the first table) vary somewhat compared to the weighted percentages (shown in the second table). Notice also how the Ns are much smaller in the unweighted run compared to the weighted run. The unweighted N refers to the actual sample size, while the weighted N approximates the size of the target population—in this case U.S. fourth-graders. For some research purposes you may wish to do unweighted analyses. For example, if you are crossing variables to form subgroups for analysis, it may be important to run unweighted frequencies to evaluate the number of cases in each cell. However, sampling weights always need to be used when you want to obtain results that are representative of the target population.

Slide 20 of 31

As just discussed, we can get correct estimates (such as percentages and mean scores) that are representative of our target population by using sampling weights. But how do we know how close our estimates are to the population values? We know that our estimates are not precise, as sampling always results in some error, or variance. Standard errors are a measure of the precision of our estimates, and the estimation of this error is called variance estimation.

Slide 21 of 31

In general, there is more uncertainty surrounding estimates derived from a complex sample than from a simple random sample of the same size. Thus, the standard errors are generally larger in estimates from complex sample designs. This is largely because of what's called a clustering effect. That is, students within the same school tend to be more similar to one another on characteristics, such as learning environment and social and economic backgrounds, than students across all schools in the population. Because of this, a simple random sample of 4,500 students drawn from across the population would be expected to give you more precise estimates for the population than a sample of 100 schools with 45 students sampled in each school. However, as explained in Common Module 3, obtaining a nationally representative sample by using simple random sampling is not practical in terms of time and cost. The clustering and multi-stage sampling design used in PIRLS, TIMSS, and PISA greatly reduces the time and cost of data collection over wide geographic areas, though the trade-off is that the estimates produced from these complex samples generally have more uncertainty

associated with them. In studies using a complex sample design, standard errors tend to get larger as sample sizes are smaller and when there is less variability among students within schools and more variability among students between schools.

Slide 22 of 31

Because studies like PIRLS, TIMSS, and PISA use a complex sample design, the formulas for calculating the standard errors are more complex than what is used for a simple random sample. Most statistical software packages assume simple random sampling. These should not be used for statistical analysis that involves hypothesis testing. They will underestimate sampling error and give incorrect p values, often indicating that differences are statistically significant when they really are not.

Fortunately, special statistical software is available, such as the IEA International Database Analyzer, which automatically uses sampling weights and correctly calculates the standard errors. Instructions on how to use and download this software are presented in the module titled ‘Considerations for Analysis of IAP Data’, or you can click on the underlined text to download it.

Slide 23 of 31

In Common Module 4 two standard error calculation procedures were discussed: Replication Techniques and Taylor Series linearization. These methods of producing standard errors use information about the sample design to calculate appropriate standard errors. Producing standard errors using SRS assumptions will lead to incorrect results. PIRLS, TIMSS and PISA use Replication Techniques, which is a method that calculates appropriate standard errors based on differences between estimates from the full sample and a series of created subsamples, or replicates. You can click on the underlined screen text, “Replication Techniques,” to go directly to Common Module 4.

PIRLS and TIMSS use a Jackknife Repeated Replication (JRR) method. There are two variables (JKZONE) and (JKREP) in the PIRLS and TIMSS data files that contain the jackknife replication information. These variables are to be used when doing analyses with PIRLS and TIMSS data in order to correctly calculate standard errors.

PISA uses the Balanced Repeated Replication (BRR) method. The actual replicate weights (a total of 80 variables) are stored in the PISA data files. These variables are to be used when doing analyses with PISA data in order to correctly calculate standard errors.

In the two slides that follow, we will take just a quick look at how the JRR and BRR methods work.

Slide 24 of 31

A major focus of PIRLS, TIMSS, and PISA is, of course, the student assessment. A challenge is to develop an assessment that comprehensively covers the subject area, or areas, without overburdening individual students. For example, at grade 8, TIMSS aims to not only provide valid and reliable measures of mathematics and science overall, but also for content domains such as algebra, data and chance, biology, and physics. Thus, large item pools are needed. However, test developers have to be sensitive to student fatigue and the fact that principals and teachers do not want the students' school day interrupted by a long assessment. Therefore, in large-scale assessments like PIRLS, TIMSS, and PISA, each student completes only a subset of the item pool.

The fact that each student completes only a subset of items means that classical test scores, such as the percent correct, are not accurate measures of student performance. Instead, using item response theory scaling (IRT), student performance in a subject can be summarized on a common scale even when different students are administered different items.

Slide 25 of 31

Using IRT, we can ask, "How would the students have performed on the test had we been able to administer all of the items to all of the students?" IRT models allow us to create a continuum on which both student performance and item difficulty can be located, linked by a probabilistic function. IRT identifies patterns of responses and uses statistical models to predict the probability of a student answering an item correctly as a function of his or her proficiency in answering other questions. It provides estimates of item parameters (such as difficulty and discrimination) that define the relationship between the item and the underlying construct measured by the test. The probability of a correct answer depends on the item parameters and ability of the examinee. Students of high ability are expected to answer both easy and difficult items correctly, while students of low ability are not expected to answer difficult items correctly. IRT model parameters are estimated for each test question, with an overall scale established as well as scales for each predefined content area specified in the assessment framework. Participating education systems contribute equally to the setting of the scales.

Slide 26 of 31

In addition to IRT, large-scale assessments like PIRLS, TIMSS, and PISA make use of multiple imputations, or what is referred to as a "plausible values" methodology. Plausible values are used to characterize scale scores for students participating in the assessment. As was mentioned earlier, to keep student burden to a minimum, students are administered a limited number of assessment items; too few to produce accurate content-related scale scores. The plausible-values methodology is used to represent what the true performance of an individual might have been, had it been observed. For each student, five random draws are generated from the estimated ability distribution of students with similar item response patterns and background characteristics. You can think of this as a regression, where the predictors are item responses and background

data. The plausible values function like point estimates of scale scores for many purposes, but they are unlike true point estimates in several respects. They differ from one another for any particular student, and the amount of difference quantifies the spread in the underlying distribution of possible scale scores for that student. While constructing plausible values, careful quality control steps ensure that the subpopulation estimates based on these plausible values are accurate. Plausible values are constructed separately for each national sample.

Slide 27 of 31

Here is a graphical presentation of five plausible values. It shows a student‘s ability distribution as estimated by IRT and 5 random draws from that distribution.

Slide 28 of 31

Here are some important points to bear in mind about plausible values.

First, always compute a statistic (such as a mean score, a correlation coefficient, or a regression coefficient) with each plausible value and **then** average the results.

Second, when using plausible values, the standard errors must be calculated correctly. The standard error is the square root of the variance, and the variance is a combination of sampling variance and measurement (or imputation) variance. Sampling variance results from selecting a subset of students and measurement variance results from administering a subset of items to each student.

As will be discussed in detail in the module titled, ‘Considerations for Analysis of IAP Data,’ when using special statistical software such as the IEA IDB Analyzer, it automatically uses plausible values correctly (including the correct computation of statistics and standard errors) when doing analyses with the achievement variables.

Finally, plausible values have been designed to provide reliable and valid achievement scores for populations (for example, average science scores of U.S. fourth-graders in TIMSS) and subpopulations (for example, average science scores of U.S. fourth-grade females in TIMSS). Plausible values should **not** be used to obtain scores for individual students.

Slide 29 of 31

Using Japan’s eighth-grade data from TIMSS 2011, here is an example illustrating the importance of calculating the correct standard errors. Regular statistical software assumes simple random sampling, and, as a result, underestimates the amount of variance in the data and gives biased standard errors. The first table shows a weighted analysis but without using the study design variables necessary for correct variance estimation, while in the second analysis the study design variables are used. Notice how the standard errors are smaller in the first analysis than in the second analysis, and notice how the first analysis **incorrectly** reports the difference in mathematics scores of girls and boys as statistically significant.

AM Statistical Software, which is free and publicly available, can be accessed by clicking on the underlined text.

Slide 30 of 31

Here is an example illustrating how plausible values are appropriate for providing estimates of achievement for a population (in this example, U.S. eighth-grade mathematics achievement from the 2011 TIMSS), but not for individuals. Notice that for individual cases there is a lot of variation across the five plausible values in the mean mathematics score. In the case highlighted, there is a difference of 63 points for two of the plausible values. That is, the plausible values do not do a good job providing an accurate and reliable measure of an individual student's mathematics achievement; they are not designed for that purpose. However, notice that for the U.S. overall there is very little variation across the five plausible values in the mean mathematics score. The difference between any two plausible values is no more than 1.32 points.

Slide 31 of 31

This module has provided you with information regarding the need for and ways to appropriately use weights, variance estimation procedures, and plausible values with data from PIRLS, TIMSS, and PISA. When analyzing these data, the procedures described within this module must be used to assure that the results are representative of the target population and that hypothesis tests are accurate. This means that weights must be applied, standard errors must be correctly calculated, and plausible values must be used while analyzing the data. These things can be done easily and automatically when using special statistical software like the IEA IDB Analyzer, which is discussed in detail in the module titled, 'Considerations for Analysis of IAP Data'.

You may now proceed to the next module in the series, or click the exit button to return to the landing page.